Wings Drugome: Analysis and extraction of the sub-workflows -- Daniel Garijo & Yolanda Gil

Date: 07/09/2011 (MM/DD/YYYY)

(Started:07/05/2011)

Analysis of the “Methods” section of the paper: Templates.

For each “Method”:

- Tools that participate. Analysis
- Use of databases/databanks. Analysis.
- Figure: how could this be a sub-workflow? (Draft aproximation)
- How this “Method” could be linked to other methods (subworkflows)?
- Other questions

For each Tool:

The analysis of the tools is mainly a:

- Feature description of the tool
- The process followed by the tool to produce the expected outcome
- What are the inputs and outputs needed for the tool to work?
- What are the parameters needed for the tool to work properly?
- Can it be run through the command line?
- Is it open source?
- Do I need to install it locally or is it a package/library that I can download and run?
- Example of usage.

For each Dataset/Databank:

- Main features of the dataset/databank
- Is it open access ?
- How do I query the dataset/ bank? (Is there a web service, or something like that?)
- Can it be accessed remotely?
- Example of query.
- Usage in the step

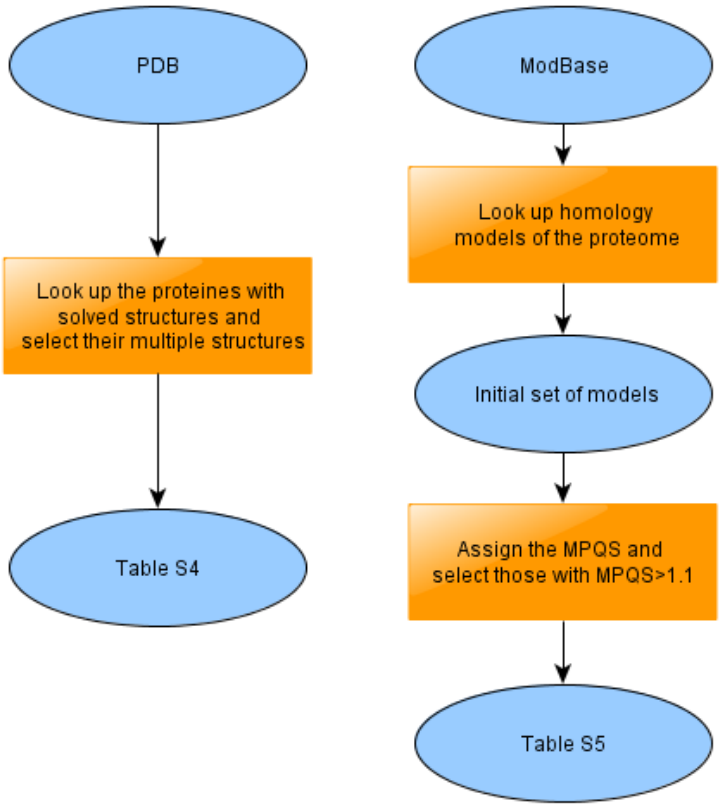
****
****
****

Structural coverage of the M.tb proteome

- Description of the method: In this step takes place the selection of the proteins and the homology models to be used in the study.
- Tools that participate. Analysis
  - **None**.
- Use of databases/databanks. Analysis.
  - **RCSB PDB**: Research Collaboratory for Structural Bioinformatics Protein Data Bank (http://www.pdb.org/pdb/home/home.do)
    - Features: The Protein Data Bank (PDB) archive is the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids.
    - Is it Open?: Yes (free of all copyright restrictions and made fully and freely available for both non-commercial and commercial use)
    - How to query it?: the website (http://www.pdb.org/pdb/search/advSearch.do), ftp server (supporting ftp and rsync access (ftp://ftp.wwpdb.org)), Web Services and RSS feeds.
    - Can it be accessed remotely? YES: http://www.pdb.org/pdb/software/rest.do
    - Query example: In the website, you can enter a protein ID and search the results. It can be also visualized in 3D with some java applets.
    - Usage in this step: Find which proteins in the M.tb proteome have solved structures in the RCSD PDB.
  - **ModBase** (homology models): http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi
    - Features: queryable database of annotated protein structure models. MODBASE contains theoretically calculated models, which may contain significant errors, not experimentally determined structures.
    - Open? Yes. Users of ModBase must cite a specific paper in their publications.
    - How to query it: ModBase Search form: http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi . Login requested for advanced features.
    - Can it be accessed remotely? (I haven't found any web services to access it)
    - Query example: (No need, since it is donde with the graphical interface)
    - Usage in this step: ModBase is used as the database where the homology models for the M.tb proteome have been retrieved. Each model is assigned a ModPipe Protein Quality Score, and if this score is greater than 1.1 then the model can be considered reliable. MPQS is a composite score comprising sequence identity to the template, coverage, and the three individual scores evalue, z-Dope and GA341. We consider a MPQS of >1.1 as reliable. A total of 1446 reliable homology models have been selected.
  - **Table S4** in the Supporting Information section:

- Information about the solved M.tb structures used in the TB-drugome. For each protein, the gene name (if available), gene accession number, protein name and corresponding PDB codes are given.. Link to the sheet: http://www.ploscompbiol.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pcbi.1000976.s009
    - **Table S5** in the Supporting Information section: Information about the M.tb homology models used in the TB-drugome. For each homology model, the ModBase model code is given, as well as the gene accession number, gene name and description of the M.tb protein. N.B. Link: http://www.ploscompbiol.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pcbi.1000976.s010

- Figure: how could this be a sub-workflow? (Draft aproximation)


PDB
Look up the proteines with solved structures and select their multiple structures
Table S4
ModBase
Look up homology models of the proteome
Initial set of models
Assign the MPQS and select those with MPQS>1.1
Table S5

Steps of the task:

1. Look up in PDB the proteins of the M.tb proteome with solved structures.
2. Select the multiple structures available for the proteins.
3. The outcome of this is table S4.
4. Look up in ModBase the homology models of the proteome.
5. Assign to each of them the MPQS.
6. Select the models with MPQS>1.1
7. The outcome of this process is table S5.

- How this “Method” could be linked to other methods (subworkflows)? The outcomes produced by this subworkflow are 2: table S4 and table S5.

- Additional questions:
    - **How is this step done?, manually?**
    - **The access to ModBase, how is it done? I just downloaded the flat files from the database and parsed them using a script**
    - **The MPQS assignation is done manually or there is an additional script/tool? The MPQS was extracted from the flat files**

Identification of FDA-approved drug binding sites

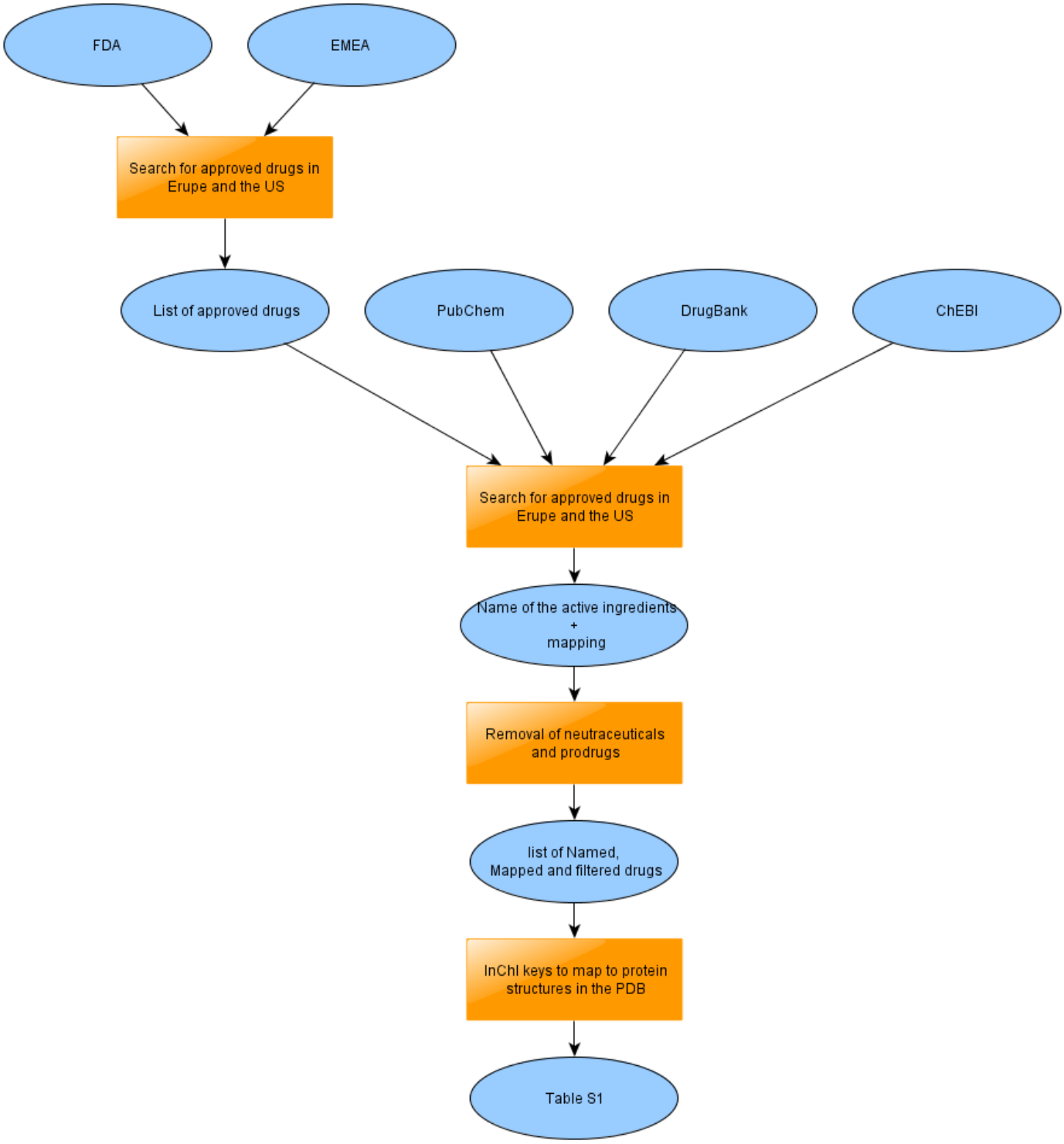
- Description of this step: In this step takes place the selection of the approved drugs binding sites in the United States and Europe used in the study.
- Tools that participate. Analysis.
    - **None**
- Use of databases/databanks. Analysis.
    - **Food And Drug Administration Orange Book**
        - Main features of the dataset/databank: http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm. Drugs for human use.
        - Is it open access? Yes
        - How do I query it? They have a Web page in http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm
        - Can it be accessed remotely? There is no need: it can be downloaded from http://www.fda.gov/Drugs/InformationOnDrugs/ucm129689.htm
        - Example of query: N/A (direct usage of the web service)
        - Usage in the step: Query to retrieve the drugs for human use in the US
    - **European Medicines Agency**
        - Main features of the dataset/databank: http://www.ema.europa.eu/ema/. Drugs for human use in Europe. Monitorized safety of the drugs in the dataset.
        - Is it open access? Yes
        - How do I query the dataset bank? Via the frontend exposed at http://www.ema.europa.eu/ema/
        - Can it be accessed remotely?There is no need. It can be downloaded from: http://www.ema.europa.eu/ema/index.jsp?curl=pages/document_library/document_listing/document_listing_000312.jsp&murl=menus/document_library/document_library.jsp&mid=WC0b01ac0580022517
        - Example of query (N/A, since it is done providing the medicine name at the web page)
        - Usage in the step: Query to retrieve the available drugs for human use in Europe.
    - **PubChem**
        - Main features of the dataset/databank: http://pubchem.ncbi.nlm.nih.gov/. Provides information on the biological activities of small molecules, and it

is organized as three diferent linked datasets (Substance, Compbound and BioAssay)
- Is it open access? Yes
- How do I query the dataset bank? Via its frontend (http://pubchem.ncbi.nlm.nih.gov/). Lists of IDs can also be queried: http://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi
- Can it be accessed remotely? ftp: ftp://ftp.ncbi.nlm.nih.gov/pubchem/ (bulk data download)
- Example of query: N/A.
- Usage in the step: Used for extraction of the names of the active ingredients of the drugs.

- **DrugBank**
  - Main features of the dataset/databank: Combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information.
  - Is it open access? Yes
  - How do I query the dataset bank? Via the web page: http://www.drugbank.ca/extractor. Results can be shown in various formats (HTML, CSV). Text query is also possible: http://www.drugbank.ca/search/advanced
  - Example of query: N/A.
  - Can it be accessed remotely?It can be downloaded: http://www.drugbank.ca/downloads
  - Usage in the step: Used for extraction of the names of the active ingredients of the drugs.
- **ChEBI**
  - Main features of the dataset/databank: http://www.ebi.ac.uk/chebi/. Stands for **Ch**emical **E**ntities of **B**iological **I**nterest (ChEBI). Centered in small molecular entity groups. It also provides an ontological classification, capturing the relationships between entity classes and their parents/childs.
  - Is it open access? Yes
  - How do I query the dataset? Via the web page: http://www.ebi.ac.uk/chebi/init.do or http://www.ebi.ac.uk/chebi/advancedSearchForward.do (advanced search)
  - Can it be accessed remotely?It can be accessed via ftp: ftp://ftp.ebi.ac.uk/pub/databases/chebi/ or downloaded from the next link: http://www.ebi.ac.uk/chebi/downloadsForward.do
  - Example of query: N/A.
  - Usage in the step: Used for extraction of the names of the active ingredients of the drugs.
- **Table S1 in the Supporting Information Section:** Information about the approved drug binding sites used in the TB-drugome. For each drug, its name,

PDB ligand code, isomeric SMILES string and known targets are listed, and the PDB codes of the protein structures with which it has been crystallized are given. Link: http://www.ploscompbiol.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pcbi.1000976.s006

- Figure: how could this be a sub-workflow? (Draft aproximation)


FDA
EMEA
Search for approved drugs in Erupe and the US
List of approved drugs
PubChem
DrugBank
ChEBI
Search for approved drugs in Erupe and the US
Name of the active ingredients + mapping
Removal of neutraceuticals and prodrugs
list of Named, Mapped and filtered drugs
InChI keys to map to protein structures in the PDB
Table S1

Steps of the task:

1. Search for drugs approved by the US and Europe in FDA and EMEA.
2. Obtain the names of the active ingredients of the drugs.
3. Map the compounds in PubChem, DrugBank and ChEBI.
4. Remove nutraceuticals and prodrugs.

5. Use InChI keys to map the remaining compunds to proteins structures in PDB, excluding non-protein crystal structures.
6. Outcome of the process is table S1.

- How this “Method” could be linked to other methods (subworkflows)? The outcome of this subworkflow is table S1. Still to be seen how it is connected to others.
- Other questions:
  - **Are the InChI keys already provided, or are they obtained using their software (http://www.iupac.org/inchi/release102.html)**

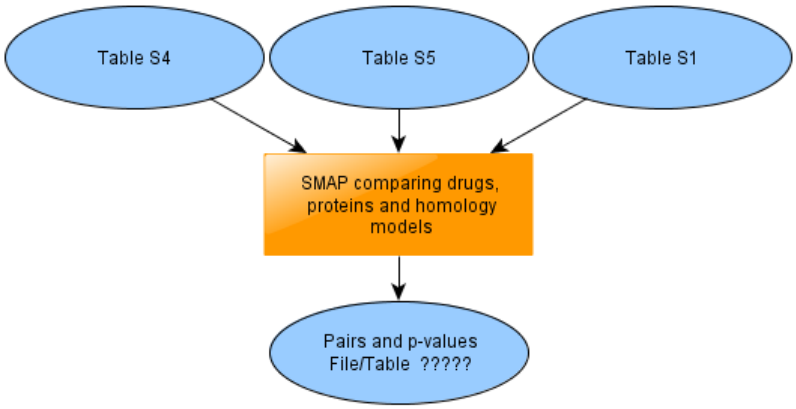
Comparison of ligand binding sites using SMAP

- Description of the step: In this step SMAP tool is used to compare the binding sites of of the 749 protein structures + 1446 homology models with the 962 binding sites of the 274 approved drugs. The outcome is a p-value for each pair compared.
- Tools that participate. Analysis
  - **SMAP**

- Features of the tool: designed for the comparison and the similarity search of protein three-dimensional motifs independent on the sequence order.
- The process followed by the tool to produce the expected outcome: Based on a sequence order independent profile-profile alignement (SOIPPA algorithm). It also uses the MWSG algorithm to align two protein structures using a maximum weighed sub-graph.
- What are the inputs and outputs needed for the tool to work? It requires 2 structures. For each it is required:
  1. PDB ID or PDB File
  2. Chain ID
- What are the parameters needed for the tool to work properly? All the parameters are specified at: http://nbcr.sdsc.edu/pub/wiki/index.php?title=SMAP_Opal_Services#Programmatic_Access
- Can it be run through the command line? It offers a WS, accesible programatically: http://nbcr.sdsc.edu/pub/wiki/index.php?title=SMAP_Opal_Services#Programmatic_Access. It can also be downloaded.
- Is it open source? Yes
- Do I need to install it locally or is it a package/library that I can download and run? I can download it from http://funsite.sdsc.edu/scb/smap/document.html#%20Installation (Installation instructions)
- Example of usage: (from the WS python GenericServiceClient.py \
  -l http://kryptonite.nbcr.net/opal2/services/SMAPPairComp \
  -r launchJob \
  -a "template_cif_chain=a template_pdb_id=1qkt query_cif_chain=a query_pdb_id=1ohp

VIRTUAL_LIGAND_RANGE_CUTOFF=5.0
LIGAND_CONTACT_DISTANCE_CUTOFF=10.0
SCORE_MATRIX=McLACHLAN
TEMPLATE_LIGAND_SITE_ONLY=false
QUERY_LIGAND_SITE_ONLY=false".
Example using the local installation of SMAP:
http://funsite.sdsc.edu/scb/smap/document.html#%20Parameter%20settings

    - SOIPPA and MWSG algorithms are not described because they are used by the SMAP tool. They are not a separate process.
- Use of databases/databanks. Analysis.
    - **None**
- Figure: how could this be a sub-workflow? (Draft aproximation)



Steps of the task:

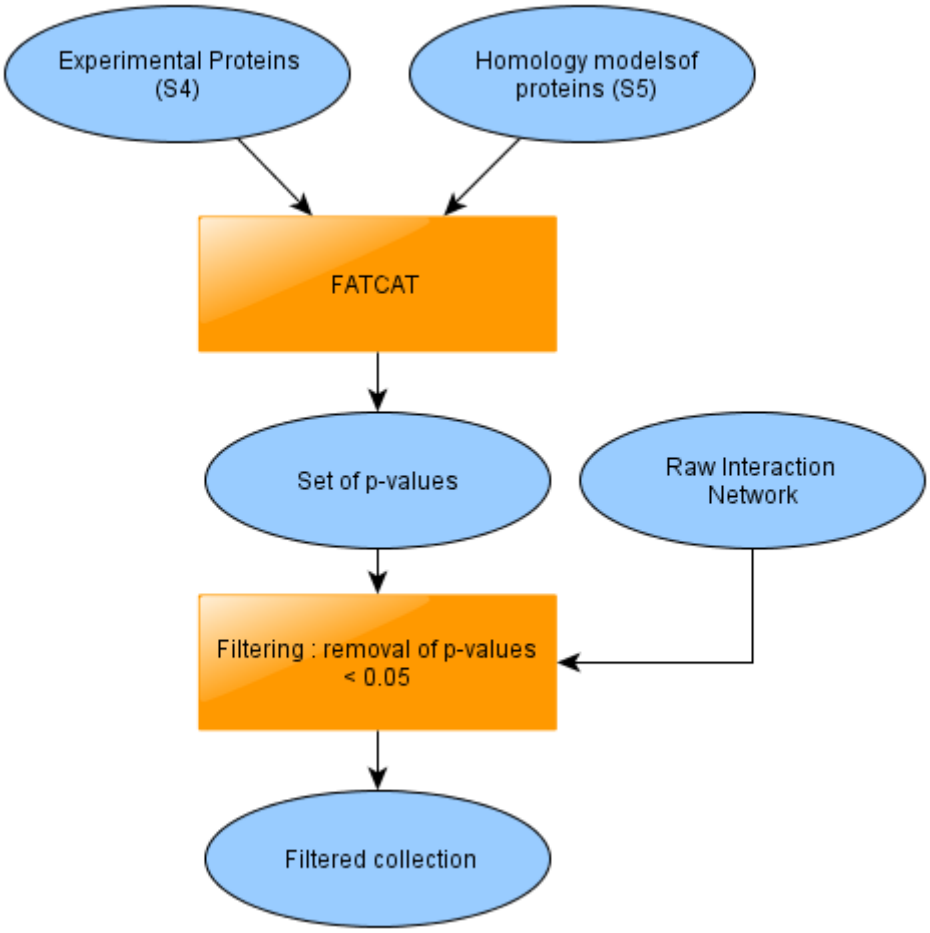
1. Takes as input table S4 (749 proteins), the 1446 homology models from table S5 as the first parameter for SMAP. The second one is the 962 bindings for the 274 drugs ~~(table S1)~~. **There is also a script to call the pairs of the proteins.**
2. Compare the binding sites with the approved drugs in an all-against-all manner.
3. For each pairwise comparison, a P-value is produced (the **p-value** is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true). This p-value represent the significance of the binding.
4. **Where is the output located? 1 FILE per each pair. Additional step to extract all the values and drop them into a table. They have a script, so we would need it. PERL scripts.**

- How this “Method” could be linked to other methods (subworkflows)? It uses the outcome of the previous 2 sub-workflows as the input to make the comparison and produce the p-values. **I assume that the output from the SMAP tool is a list of p-value result, one for each comparison (this is not specified in the paper). It would be a list of 2195*962 = 2.111.590 entries (or a table of 2195 rows and 962 columns)**

Comparison of global protein structures using FATCAT

- Description of the step: In this step takes place the use of FATCAT
- Tools that participate. Analysis
    - **FATCAT**
        - Feature description of the tool: Flexible structure AlignemenT by Chaining Aligned fragment pairs allowing Twists. It is used to report the overall similarity between 2 structures, using a p-value to measure it. P-value less than 0.05 means that they are similar.
        - The process followed by the tool to produce the expected outcome. The user fills a form providing either the PDB code, file in PDB format or SCOP domain code of the two resources to be aligned and submits it to the server. The response is a p-value with the similarity and some additional outcomes from the server. Example: http://fatcat.burnham.org/fatcat/examples/1ufhA_1gheA/. Form available at: http://fatcat.burnham.org/fatcat-cgi/cgi/fatcat.pl?-func=pairwise
        - What are the inputs and outputs needed for the tool to work? PDB code, file in PDB format or SCOP domain code of the 2 entites to compare.
        - What are the parameters needed for the tool to work properly? No additional parameters are required.
        - Can it be run through the command line?Apparently **Not**. But it looks like someone just could fill the form automatically, submit it and treat the results.
        - Is it open source? Yes
        - Do I need to install it locally or is it a package/library that I can download and run? Not available to install it locally
- Use of databases/databanks. Analysis.
    - **None**

- Figure: how could this be a sub-workflow? (Draft aproximation)



Steps:

1. Query the PDB for the PDB files. **Are these the same PDB files as in table S4?**
2. Use FATCAT to filter the non-similar structures. Assuming it is table s4, Is each PDB file compared to all the other ones?
3. The output would be (assuming that we are taking the 749 structures from the PDB in table S4) a file with 794*794 = 630.436 p-values
4. Remove the pairs with high similarity (pvalue<0.05)

- How this “Method” could be linked to other methods (subworkflows)?
  - It is not clear how this relates to other sub workflows. This step occurs before the SMAP step, which uses as input the table S1? How is the list with the non similar structures used later?
- Other questions.
  - **Should ask if they do this step manually or if they access some other service to send multiple queries. Check with Sarah - I used Andreas’ JFatCat program (Java implementation of FATCAT)**
  - **The PDB files mentioned in the paper, is it a reference to table s4? if not, which other PDB files are they referring to in this step?**

Visualization of the protein-drug interaction network

- Description of the step: This step describes the use of yEd as a Graphical editor to visualize the protein-drug interaction.
- Tools that participate. Analysis
    - **yEd Graph Editor**.
        - Feature description of the tool: Diagram editor.
        - The process followed by the tool to produce the expected outcome. N/A
        - What are the inputs and outputs needed for the tool to work? **The network from FATCAT. A Formatting is required in order to represent the network with yEd. Cytoscape could be an alternative. Check with Lei and Sarah if they have it. I wrote a script to generate the input for yEd in the correct format.**
        - What are the parameters needed for the tool to work properly?N/A
        - Can it be run through the command line?Yed uses a graphical interface. It can not be run directly with the command line. In fact, it is not allowed: **Clarifies that using yEd as part of an automated process is not allowed.**
        - Is it open source? It is free, but not open source.
        - Do I need to install it locally or is it a package/library that I can download and run? You need to install it locally
        - Example of usage. N/A
- Use of databases/databanks. Analysis.
    - **NCBI Entrez**:
        - Main features of the dataset/databank: collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB
        - Is it open access ? YES
        - How do I query the dataset/ bank? (Is there a web service, or something like that?) It offers a web access to the database, filling a form or by key words: http://www.ncbi.nlm.nih.gov/books/NBK44863/. FTP: ftp://ftp.ncbi.nlm.nih.gov/genbank/, ftp://ftp.ncbi.nih.gov/refseq/
        - **Can it be accessed remotely? I haven't found any other remote access. MANUALLY by filling a form. Maybe we can automatize this writing a component.**
        - Example of query. N/A
        - Usage in the step. To query the names of the M.tb proteins, in order to avoid inconsistences with the naming proteins of the PDB.
- Figure: how could this be a sub-workflow? (Draft aproximation)
    - **No need to represent the workflow for this task (I think)**

Steps:

1. (Assumption) Treat the information produced by SMAP into a table/Matrix so it can be read directly by yEd.

2. Query the M.tb protein names to the Entres protein database.
3. Represent the graph using yEd.

- How this “Method” could be linked to other methods (subworkflows)?
  - The input used for the graph representation is the same one produced by SMAP.
- Other questions:
  - What is the input received by yEd? (I have assumed it is the one from SMAP)
  - Is this process necessary for the study or for just seeing the results ?
  - How is this step linked to other parts of the workflow?

Flux balance analysis

- Description of the step: In this step both the in vivo essentiality and the in vitro essentiality are calculated using different toolboxes and databases. (I assume that non essential genes are discarded).
- Tools that participate. Analysis
  - **COBRA Toolbox:**
    - Feature description of the tool. http://opencobra.sourceforge.net/openCOBRA/Welcome.html. Constraints Based Reconstruction and Analysis (COBRA) focuses on employing physicochemical constraints to define the set of feasible states for a biological network in a given condition based on current knowledge.
    - The process followed by the tool to produce the expected outcome. It is used to delete single genes on the iNJ661 model.
    - What are the inputs and outputs needed for the tool to work? It uses the iNJ661 model grown in 7H9 media. **No additional information is provided. How is the iNJ661 used? which functions/scripts of the COBRA Toolbox are used?**
    - What are the parameters needed for the tool to work properly? Installations instructions+ usage available at : http://www.nature.com/protocolexchange/protocols/2097#/procedure
    - Can it be run through the command line? Yes, because they are Matlab scripts.
    - Is it open source? NO. It is designed to run with Matlab, which is not open source. However the toolbox is free.
    - Do I need to install it locally or is it a package/library that I can download and run? Local install (Matlab+Cobra toolbox download)
    - Example of usage. N/A, since the methods of the toolbox used for the study are not known.
- Use of databases/databanks. Analysis.
  - **GSMN-TB**
    - Main features of the dataset/databank: Web-based genome -scale network model used to carry out the flux balance analysis.
    - Is it open access ?Yes
    - How do I query the dataset/ bank? (Is there a web service, or something like that?) http://sysbio3.fhms.surrey.ac.uk/cgi-

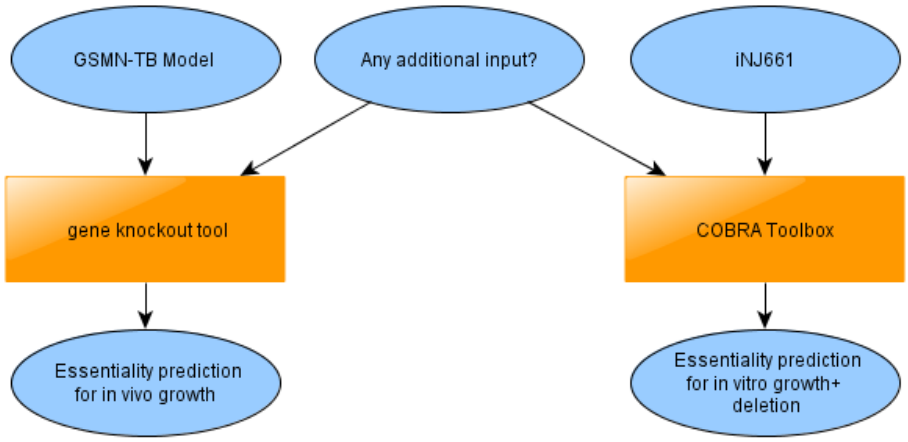
bin/fba/fbapy?model=GSMN_TB-Viv&cmd=methods. A web form based query is the only available way to use the service. It can be modified: http://sysbio3.fhms.surrey.ac.uk/cgi-bin/fba/fbapy?model=GSMN_TB-Viv&cmd=fba.

- Can it be accessed remotely?Aparently not. **Have the authors accessed in some other direct way? No, it was just accessed using the web. This didn't matter, as it was only for a small number of genes.**
- Example of query. http://sysbio3.fhms.surrey.ac.uk/cgi-bin/fba/fbapy?model=GSMN_TB-Viv&cmd=fba already shows a query. It returns a file like this, showing the FBA: http://sysbio3.fhms.surrey.ac.uk/cgi-bin/fba/fbapy?sid=sid5008843&cmd=showfba&model=GSMN_TB-Viv
- Usage in the step. This model is used to calculate the essentiality prediction under conditions optimized for *in vivo* growth.

  - **iNJ661**

- Main features of the dataset/databank
- Is it open access ? Yes
- How do I query the dataset/ bank? (Is there a web service, or something like that?) It can be downloaded through the next link: http://www.biomedcentral.com/1752-0509/1/26/additional
- Can it be accessed remotely? There is no need, it can be downloaded as a .xls sheet
- Example of query.N/A
- Usage in the step. Used to calculate the *in vitro* essentiality (with the COBRA Toolbox).

- Figure: how could this be a sub-workflow? (Draft aproximation)



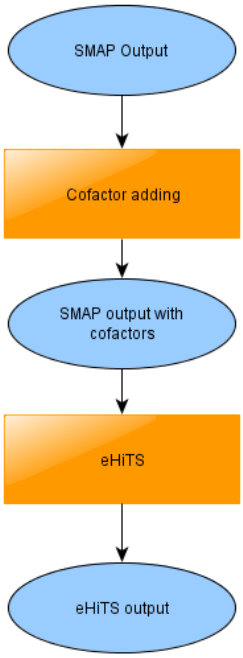
Steps :

1. Use GSMN-Tb to carry out the FBA computations.
2. Use the single gene knockout tool to run essentiality prediction.
3. Constrain some genes in order to simulate multiple gene knockouts.
4. Use iNJ661 as input to the COBRA Toolbox to perform single gene deletions and determine in vitro essentiality.

- How this “Method” could be linked to other methods (subworkflows)? **??????? (it is defined as an independent step)**
- Other questions

Molecular docking using eHiTS

- Description of the step: in this step, molecular docking to predict the binding pose and affinity of the drug molecule to the drug proteine takes place, using eHiTS.
- Tools that participate. Analysis
    - **New tool for doing the docking: Autodock vina. They have scripts for running it.**
    - **eHiTS Lightning**
        - Feature description of the tool. http://www.simbiosys.com/ehits/ehits_benefits.html (According to the web page): Fast, accurate, full automated, customizable tool used for docking studies.
        - The process followed by the tool to produce the expected outcome. Takes the input produced by the SMAP tool. For those proteins with cofactors, the cofactor was added as the last residue in the protein structure prior to docking. Nothing is said about the outcome produced by the tool.
        - What are the inputs and outputs needed for the tool to work? (Assumption) The significance value of the pairs analyzed with the SMAP tool.
        - What are the parameters needed for the tool to work properly? **??? (I had no access to the user manual)**
        - Can it be run through the command line? **???? (ask the authors of the study/ ask for a demo) Yes e.g.: ./ehits.sh -receptor receptor_file.pdb -clip clip_file -ligand ligand_file.pdb -workdir . -accuracy 6 -out output_file.sdf**
        - Is it open source?No, you have to request for a demo and then purchase the product.
        - Do I need to install it locally or is it a package/library that I can download and run? Local install.
        - Example of usage. N/A (there is no documentation without a registration)
- Use of databases/databanks. Analysis.
    - **None**
- Figure: how could this be a sub-workflow? (Draft aproximation)



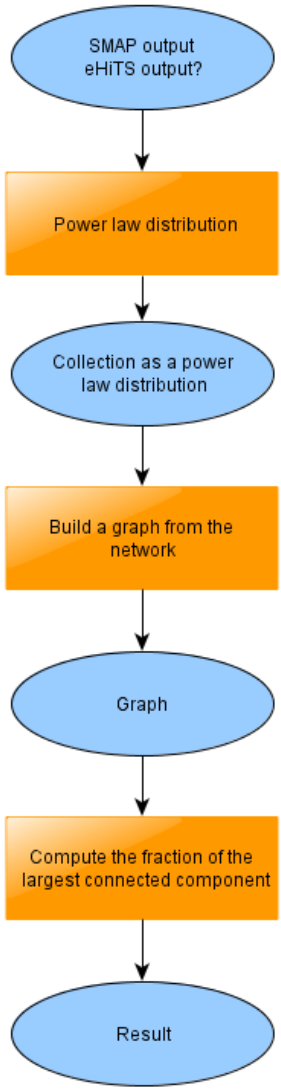
Steps:

1. Take the output produced by SMAP.
2. For those proteins with cofactors, add the cofactor as the last residue in the protein.
3. Parameters: search space of 10A^3, accuracy level = 6.
4. eHiTS outcome is produced as a **??? SDF file with the resulting conformations of the molecule and their corresponding energy scores**

- How this “Method” could be linked to other methods (subworkflows)? Takes as input the output produced by SMAP. **Nothing is said about the outcome produced.**
- Other questions
  - **Ask the authors about the eHiTS tool. Is it used automatically? What is the format of the outcome? Are there any additional parameters?**
  - **Are there any additional processes necessary to use the tool? Postprocessing? After running eHiTS using the command above, the conformation with the best score (and its score) was extracted using a script.**

Network analysis

- Description of the step: Consruction of a drug-target network protein graph.
- Tools that participate. Analysis
  - **None**
- Use of databases/databanks. Analysis.
  - **None**
- Figure: how could this be a sub-workflow? (Draft aproximation)



Steps:

1. Get the output from SMAP step or eHiTS step.
2. Fit the number of targets and their connectivity to a power law distribution.
3. Build a graph from the drug target network.
4. Compute the fraction of the largest connected component by dividing the number of proteins in the largest single linkage cluster by the total number of proteins in the graph.

- How this “Method” could be linked to other methods (subworkflows)? The input is either obtained from the SMAP step or the eHiTS step. nothing is said about the output produced in this step.
- Other questions
  - **Is this step produced manually? Ask Lei about this**
  - **Is the input taken from the SMAP setp, the eHiTS step or other steps?**
  - **What is the format of the input produced. Is it stored somewhere in a table? Format?**

- **Additional tools used?**
- **What is the format of the output?**

Hierarchical clustering of protein and drug binding profiles

- Description of the step: Hierarchical cluster of the protein and drug binding profiles, using GenePattern 2.0.
- Tools that participate. Analysis
  - **GenePattern 2.0:**
    - Feature description of the tool: genomic analysis platform that provides access to more than 150 tools for gene expression analysis, proteomics, SNP analysis, flow cytometry, RNA-seq analysis, and common data processing tasks.
    - The process followed by the tool to produce the expected outcome. I assume it takes the result from eHiTS or SMAP as input file. What are the inputs and outputs needed for the tool to work?
    - What are the parameters needed for the tool to work properly? The parameter is the city block distance.
    - Can it be run through the command line? It looks that it can be directly accessed from Java and Matlab: http://www.broadinstitute.org/cancer/software/genepattern/tutorial/gp_programmer.html?Matlab_doc#_Using_GenePattern_from_Java. Libraries available at: http://genepattern.broadinstitute.org/gp/pages/downloadProgrammingLibaries.jsf.
    - Is it open source? It is free, but I believe is not open.
    - Do I need to install it locally or is it a package/library that I can download and run? You can either http://www.broadinstitute.org/cancer/software/genepattern/installer/latest/install.htm (Setting up your own server), or use the website (login required) http://genepattern.broadinstitute.org/gp/pages/index.jsf
    - Example of usage: *JobResult result = gpClient.runAnalysis("urn:lsid:broad.mit.edu:cancer.software.genepattern.module.analysis:00009:5", new Parameter[]{new Parameter("input.filename", ""), new Parameter("column.distance.measure", "2"), new Parameter("row.distance.measure", "0"), new Parameter("clustering.method", "m"), new Parameter("log.transform", ""), new Parameter("row.center", ""), new Parameter("row.normalize", ""), new Parameter("column.center", ""), new Parameter("column.normalize", ""), new Parameter("output.base.name", "<input.filename_basename>")});* **(Java)**
- Use of databases/databanks. Analysis.
  - **None**
- Figure: how could this be a sub-workflow? (Draft aproximation)
  - **There is no need, since it is only one step.**

Steps:

1. Take the input from SMAP/eHiTS
2. Use the clustering analysis to make a hierarchy.
3. Nothing is said about the output.

- How this “Method” could be linked to other methods (subworkflows)?
- Other questions

Comparison of drug chemical similarity

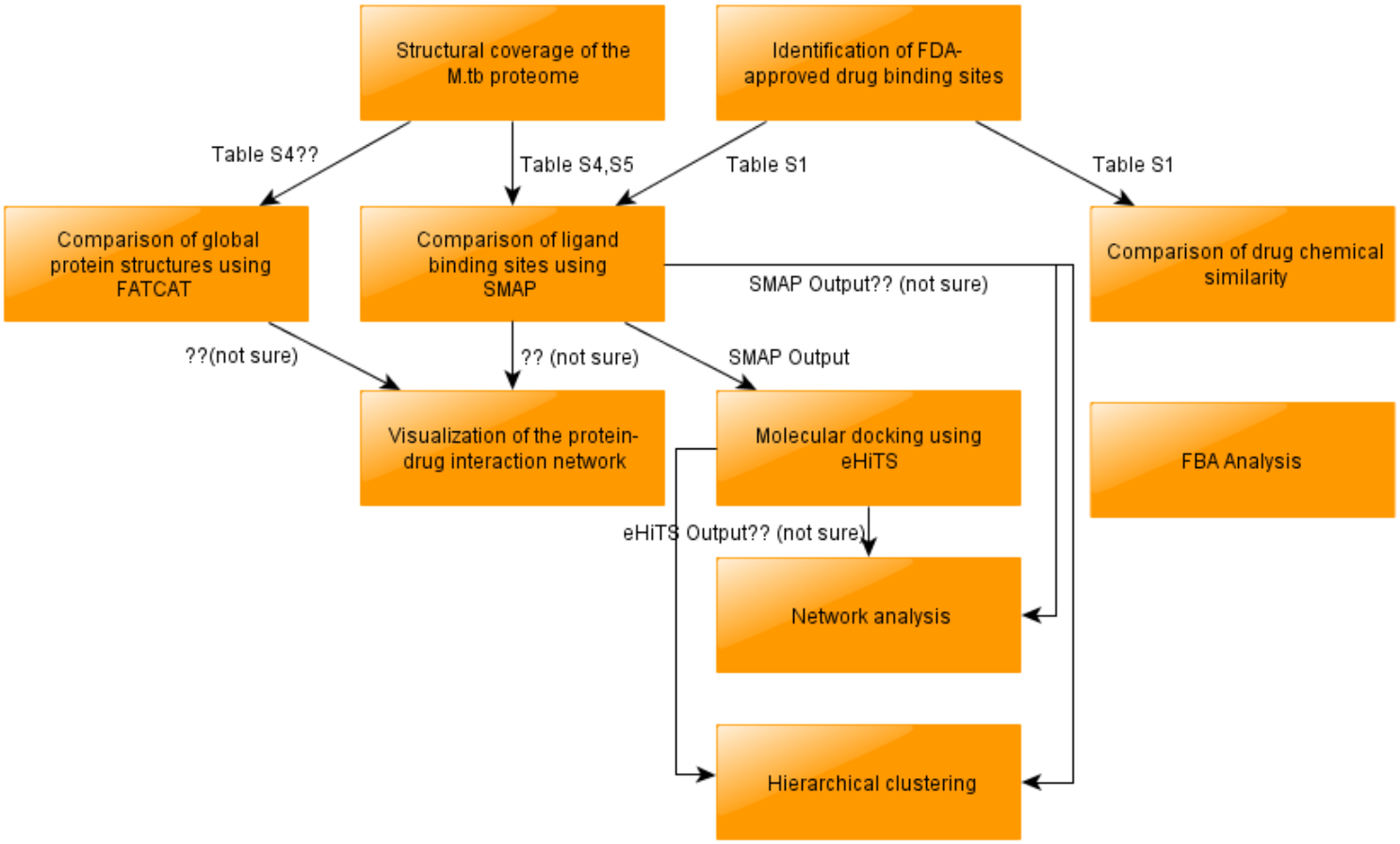
- Description of the step: 2D fingerprint similarity.
- Tools that participate. Analysis
  - **OpenBabel 2.1.1:**
    - Feature description of the tool: http://openbabel.org/wiki/Open_Babel_2.1.1 chemical toolbox designed to speak the many languages of chemical data.
    - The process followed by the tool to produce the expected outcome. It gets table S1 (drugs) (**assumption**) as input to calculate the similarity. **Nothing is said about the output. Maybe is it used for another previous step?**
    - What are the inputs and outputs needed for the tool to work? Table S1.
    - What are the parameters needed for the tool to work properly?
    - Can it be run through the command line?Yes: The **obabel** command line program converts chemical objects (currently molecules or reactions) from one file format to another. The GUI interface is an alternative to using the command line and has the same capabilities.
    - Is it open source? Yes
    - Do I need to install it locally or is it a package/library that I can download and run? Libraries can be used from Java: http://openbabel.org/docs/2.3.0/UseTheLibrary/Java.html
    - Example of usage. The link above covers them. They can be found in the tutorials too: http://openbabel.org/wiki/Tutorial:Fingerprints
      Eg: PROMPT> babel -L fingerprints
    -
- Use of databases/databanks. Analysis.
  - **None**
- Figure: how could this be a sub-workflow? (Draft aproximation)
  - **There is no need, since it is just one step**

  Steps:
  1. Input: Table S1.
  2. Process: OpenBabel
  3. Output: **????**
- How this “Method” could be linked to other methods (subworkflows)? **Nothing is said in the step**
- Other questions
  - **How is the output of this method used for other methods?**

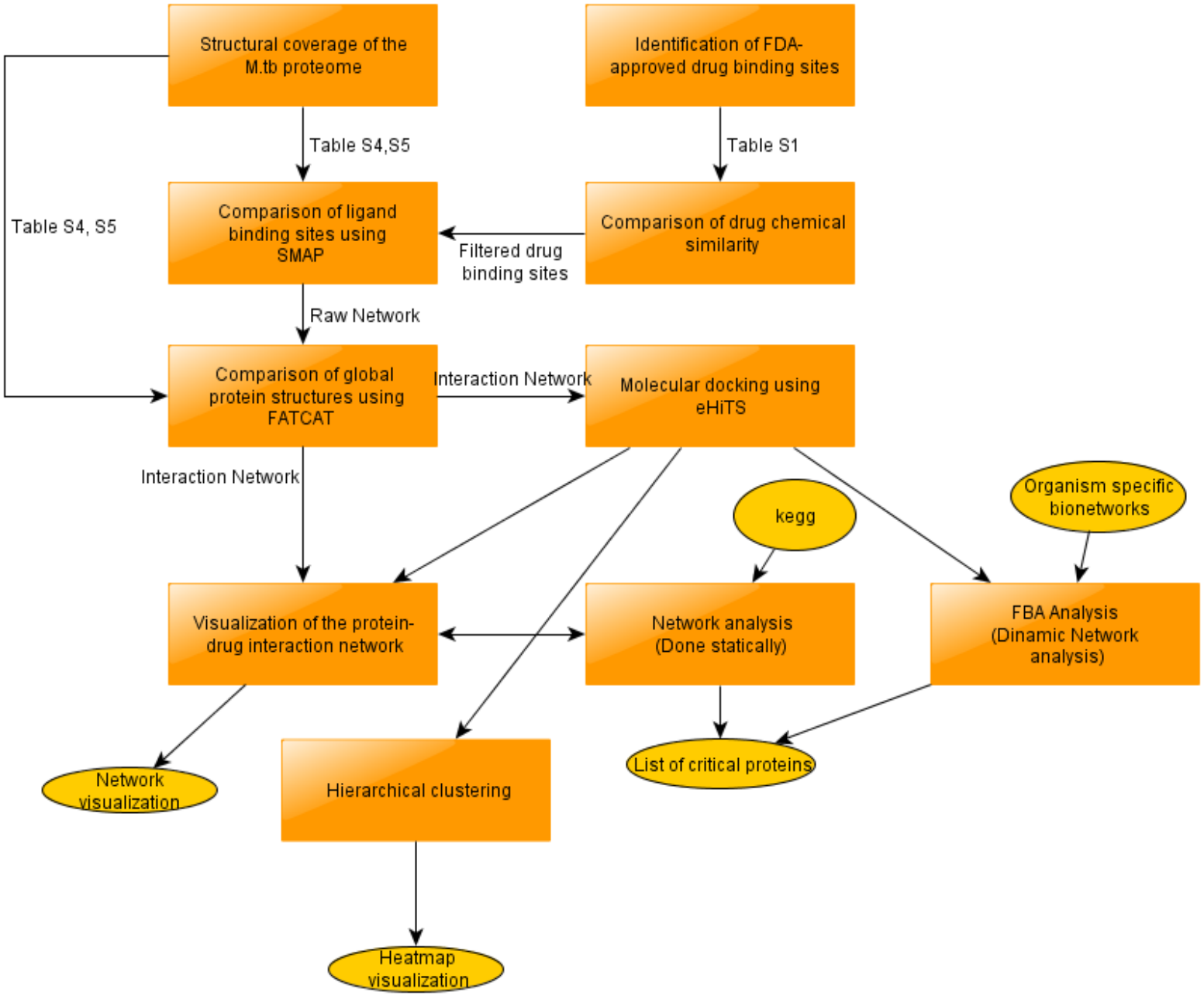
Additional steps/files:

http://funsite.sdsc.edu/drugome/TB/#Summary

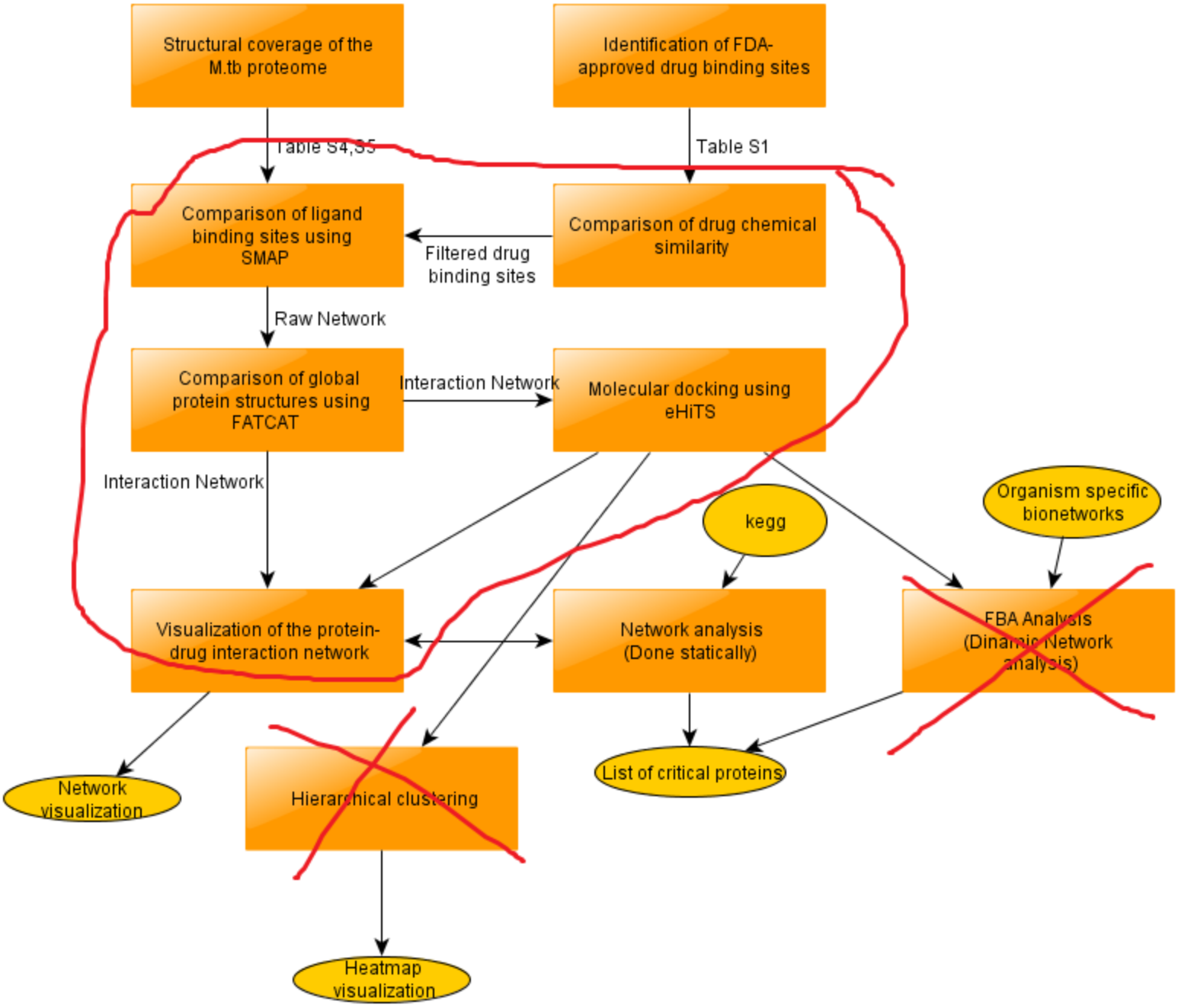
Overview: connection between all the methods



Overview: connection between all the methods (FIX)


Structural coverage of the M.tb proteome
Identification of FDA-approved drug binding sites
Table S4,S5
Table S1
Table S4, S5
Comparison of ligand binding sites using SMAP
Comparison of drug chemical similarity
Filtered drug binding sites
Raw Network
Comparison of global protein structures using FATCAT
Interaction Network
Molecular docking using eHiTS
Interaction Network
kegg
Organism specific bionetworks
Visualization of the protein-drug interaction network
Network analysis (Done statically)
FBA Analysis (Dinamic Network analysis)
Network visualization
Hierarchical clustering
List of critical proteins
Heatmap visualization

FIRST STEPS:


Structural coverage of the M.tb proteome
Identification of FDA-approved drug binding sites
Table S4,S5
Table S1
Comparison of ligand binding sites using SMAP
Comparison of drug chemical similarity
Filtered drug binding sites
Raw Network
Comparison of global protein structures using FATCAT
Interaction Network
Molecular docking using eHiTS
Interaction Network
kegg
Organism specific bionetworks
Visualization of the protein-drug interaction network
Network analysis (Done statically)
FBA Analysis (Dinamic Network analysis)
Network visualization
Hierarchical clustering
List of critical proteins
Heatmap visualization